Claim Amendments

This listing of claims replaces all prior versions and listings of claims in the application.

Claims 1 – 118 (Canceled).

119. (New) A method for automating the extraction of information from a semi-structured document characterized by a document type that comprises design and structural characteristics of a set of similar documents, the method comprising: designing a target extraction template for the terms of the document type; supporting the creation of a control set of documents containing the terms manually tagged to the extraction template; automatically generating a skeleton of extraction model tree for every term; training the models by automatically optimizing selectors of the term extraction models to the best compliance with the control set tagging; and using the optimized model to automatically extract information from the document.

120. (New) The method of claim 119, further comprising using specialized invariants to select generic components of information from the document.

121. (New) The method of claim 119, further comprising tracking and analyzing changes made to initially extracted information and subsequent re-optimization of models.

122. (New) The method of claim 119, further comprising analyzing an additional semi-structured document and updating the model selectors or its structure if a change in accuracy of the term extraction model exceeds a threshold.

123. (New) The method of claim 119, further comprising: (a) retaining specific information about a set of semi-structured documents to serve as a template for new semi-structured document introduction; (b) comparing any new semi-structured document with a pattern represented by specific information known to be suitable for searching for text based on the retained specific information about

the set of semi-structured documents; (c) assessing if the result of (b) is within a threshold of the result of (a).

124. (New) The method of claim 123, as applied to knowledge that a given company employs similar patterns for subsequent versions of similar documents identifying the company to which the documents pertain.

125. (New) The method of claim 119, in which terms can be assigned a term class for at least one of immediate validation, synonym support, and vocabulary management.

126. (New) The method of claim 119, further comprising automatically comparing first and second extracted data to each other to identify extraction errors.

127. (New) A method of manually tagging and extracting terms from a semi-structured document while automatically collecting key indicators for pattern recognition, in which the tagging is the sole generation point of statistics needed for creation and optimization of an extraction model.

128. (New) A method of using an extraction template having terms to extract data from a semi-structured document having tagged values, comprising providing at least one of: a many-to-many relationship between the tagged values and the terms in the extraction template; a many-to-one relationship between the tagged values and a single term; or a one-to-may relationship between a single tagged value and a plurality of multiple terms.

129. (New) A method of extracting data from a semi-structured document having a source format, comprising providing a generalized spatial and contextual file format that is independent of the source format.

130. (New) The method of claim 129, in which the generalized spatial and contextual file format specifies at least one of context on the document, page, table, row, column, and offset.

131. (New) The method of claim 129, in which the semi-structured document is an EDGAR electronic filing and the method further comprises providing at least

one of access, navigation, selection, downloading, conversion into the generalized format, and insertion into a document repository.

132. (New) The method of claim 129, in which the semi-structured document is in a format selected from the group consisting of PDF, HTML, and text, and the method further comprises providing at least one of access, navigation, selection, downloading, conversion into the generalized format, and insertion into a document repository.

133. (New) A method of extracting data from a semi-structured source document, comprising providing source links for extracted data at a term level without modifying the source document, and further in which reference to the source document is provided through an abstraction enabled by a generalized intermediate format.

134. (New) A method of quality control in a process of collecting data from a semi-structured source document, comprising providing at least one of document-type specific controls; system-wide controls; automated data cross-checks; and manual quality assurance measures.

135. (New) The method of claim 134, in which the document-type specific controls are applied to the extracted content and include at least one of validation of specific data types, application of pre-assigned values, referencing of synonym lists, and application of user-defined validation rules.

136. (New) The method of claim 134, in which providing automated data cross-checks comprises automatically cross-checking currently extracted data against previously extracted data to identify potential data extraction errors.